em questão
E-ISSN 1808-5245

Integração de bases de dados em estudos bibliométricos: a produção científica nacional em Zika vírus

Maria Simone de Menezes Alencar

Doutora; Universidade Federal do Estado do Rio de Janeiro, Rio de Janeiro, RJ, Brasil;
simone.alencar@unirio.br

Rosany Bochner

Doutora; Fundação Oswaldo Cruz, Rio de Janeiro, RJ, Brasil;
rosany.bochner@icict.fiocruz.br

Daniel Giacometti Amaral

Mestre; NanoBusiness Informação e Inovação Ltda, Rio de Janeiro, RJ, Brasil;
dgiacometti.amaral@gmail.com

Resumo: A partir do estudo de caso da produção científica brasileira envolvendo o vírus Zika, este artigo tem como objetivo discutir a importância da integração de dados de diferentes fontes para melhor representar o universo da pesquisa, destacando como a restrição de fontes impacta nos resultados de análises bibliométricas. Para identificação da produção brasileira na área por meio da seleção de artigos de periódicos com pelo menos um autor brasileiro, foram utilizadas quatro bases informacionais: duas fontes gerais, a Web of Science e a Scopus; uma fonte específica da área de estudo, o PubMed; e uma fonte de pesquisa de informação nacional, a SciELO. A análise dos resultados indica a relevância da integração de diferentes bases de dados em estudos bibliométricos como meio de minimizar distorções e fornecer uma visão consistente do universo da pesquisa científica em determinado tema.

Palavras-chave*:* Bibliometria. Bases de dados. Zika. Fontes de informação.

1 Introdução

Os estudos para avaliação da produção científica baseados em bibliometria usualmente utilizam apenas uma base de dados de artigos científicos, de acordo com a disponibilidade de metadados que a fonte possui e o objetivo do estudo. A fonte mais utilizada para esse tipo de estudo é a Web of Science (WoS), devido

à sua amplitude, cobertura e completude em relação aos metadados (MUELLER, 2013).

No entanto, várias pesquisas têm sido feitas para avaliar resultados utilizando outras fontes de informação, comparando características gerais e cobertura de bases. Como exemplo, pode-se citar Jacsó (2005), Burnham (2006), Norris e Oppenheim (2007), Falagas, Pitsouni, Malietzis e Pappas (2008), Gavel e Iselid (2008), Neuhaus e Daniel (2008), Archambault et al., 2009; Harzing e Alakangas (2016) e Mongeon e Paul-Hus (2016).

Um aspecto importante em relação à base de dados refere-se aos idiomas dos artigos. Van Leewen et al. (2001) destacam que o valor dos indicadores de impacto das atividades de pesquisa no nível de uma instituição ou país depende fortemente de se incluir ou excluir publicações em periódicos cobertos pela WoS escritos em outros idiomas além do inglês. Ao analisarem essa temática, Moed, Markusova e Akoev (2018) identificaram que as contagens de publicações russas dependem fortemente do banco de dados utilizado, reforçando o potencial impacto limitante do idioma nos estudos bibliométricos. Dessa forma, os autores ressaltam que o uso de indicadores derivados da WoS e, especialmente, da Scopus, como ferramentas na mensuração do desempenho da pesquisa e orientação internacional do sistema científico russo, deve ser utilizado com ressalvas. Ainda que tenham se limitado a análise do contexto russo, as conclusões de Moed, Markusova e Akoev (2018) indicam uma realidade compartilhada por outros países.

Em setembro de 2014, pesquisadores que participavam da 19° Conferência Internacional de Indicadores em Ciência e Tecnologia (STI-2014), em Leiden, Holanda, propuseram um conjunto de dez recomendações para orientar sobre o uso de métricas em avaliação da pesquisa, documento este chamado de Manifesto de Leiden (2015). A terceira recomendação refere-se à importância de proteger a excelência da pesquisa desenvolvida e publicada localmente. Em geral, os estudos bibliométricos tomam como base as publicações em inglês, mas em muitas áreas há inúmeras pesquisas que apresentam dimensão regional ou nacional em que o idioma inglês não é o mais adequado. O pluralismo da sociedade tende a ser suprimido para criação de

artigos de alto impacto, em periódicos de língua inglesa. Métricas desenvolvidas a partir de literatura de alta qualidade em outros idiomas são úteis para identificar e premiar a excelência na pesquisa relevante localmente.

Um exemplo disso pode ser observado nos estudos nacionais recentes acerca da produção científica envolvendo o vírus Zika, quando pesquisadores decidiram adotar uma única base para realizar suas buscas, sendo que essa não foi a mesma para todos os trabalhos. Dentre os seis trabalhos analisados, três bases foram apontadas, Scopus, WoS e PubMed. Araújo, Silva e Guimarães (2016), Ribeiro et al. (2016) e Araújo et al. (2017) adotaram a base Scopus; Gheno et al. (2016, 2017) optaram pelo uso da WoS; Nunes et al. (2016) utilizaram a base de dados do PubMed. Analisando os resultados desses trabalhos foi possível verificar que a base Scopus se mostrou mais abrangente do que a WoS. No entanto, sabe-se que há trabalhos indexados na WoS que não estão na Scopus e vice-versa, assim como há trabalhos na PubMed, ou mesmo na SciELO, que não estão indexados nas outras bases.

Refletindo sobre esses achados, e na busca por uma pesquisa mais abrangente de produção científica, Alencar, Bochner e Giacometti (2018) apresentaram avanços metodológicos sobre a integração de bases de dados e discutiram o comparativo da posição dos autores mais produtivos dentre as bases do estudo no 6° Encontro Brasileiro de Bibliometria e Cientometria (6° EBBC). Aprofundando essa investigação, o objetivo deste estudo é discutir a importância da integração de dados de diferentes fontes para melhor representar o universo da pesquisa, destacando como a restrição de fontes impacta nos resultados de análises bibliométricas.

Como objeto de estudo foi escolhida a temática do Zika vírus, eleito pela Organização Mundial da Saúde (OMS) em 1º de fevereiro de 2016 como uma emergência de importância internacional (BUENO et al., 2017). O estudo tem como foco a produção científica brasileira envolvendo o tema a partir da seleção de artigos de periódicos com pelo menos um autor brasileiro indexados em diferentes bases de dados.

2 Metodologia

A metodologia proposta para desenvolvimento do presente estudo foi estruturada em três etapas centrais, sendo a primeira referente a coleta de dados, a segunda ao tratamento de dados e integração das bases e a terceira à análise dos dados. O detalhamento dessas etapas é apresentado a seguir.

2.1 Coleta de dados

Para a coleta de dados referente a produção científica envolvendo o Zika vírus foram utilizadas quatro bases informacionais: duas fontes gerais, a WoS e a Scopus; uma fonte específica da área de estudo, o PubMed; e uma fonte de pesquisa de informação nacional, a SciELO.

A estratégia utilizada em todas as bases foi o uso dos termos "zika", "zikv" e "zikav" no campo correspondente a título, resumo e palavras-chave, abrangendo publicações no período de 2014 a 2018. Nas quatro fontes informacionais, os registros foram refinados de forma a recuperar apenas artigos de periódicos, com pelo menos um dos autores com instituição de origem localizada no Brasil. Além disso, cerca de 5% de artigos coletados em cada fonte foi avaliado por meio de leitura para identificar sua pertinência e validar o conjunto de dados.

2.2 Tratamento de dados e integração das bases

Os dados coletados foram tratados com o auxílio do software *VantagePoint*[1] para eliminação de duplicatas, realização de refinamentos, padronização de campos e integração dos conjuntos de dados das diferentes bases utilizadas.

Inicialmente, foram eliminadas duplicatas presentes no conjunto de dados da própria base. A duplicidade de registros pode ocorrer devido a inconsistências na indexação como, por exemplo, o mesmo artigo indexado como *preprint* e publicado.

Em seguida, foram identificados campos em comum entre as bases que pudessem ser integrados, tendo em vista as diferentes estruturas de indexação utilizadas. Durante a avaliação das informações disponíveis nas quatro bases, foram identificados os seguintes campos em comum para composição da análise: título do artigo, autores, periódico, ano de publicação e países dos autores. Foram também utilizados campos para facilitar identificação de duplicatas entre as bases, tais como DOI, volume, número e página inicial. O campo de instituição do autor não pôde ser utilizado, pois as bases Scopus e SciELO só apresentam o endereço completo do autor, conforme consta no artigo original, não o delimitando em subcampos como instituição, endereço e país.

Foi observada ainda a necessidade de padronização de alguns campos para harmonização entre as bases, sendo eles: título do artigo, nome do periódico, país e autor. Como exemplo, pode-se citar o fato dos títulos dos artigos oriundos da base PubMed possuírem pontuação no final, sendo necessário seu processamento para exclusão, bem como a necessidade de padronizar nomes de países na base SciELO, envolvendo casos como o Brasil, para o qual pode ser encontrada a grafia em língua inglesa (Brazil). Os nomes de periódicos também estão registrados de diferentes formas nas bases: no PubMed, eles aparecem abreviados, e na WoS, por extenso, por exemplo. No que se refere aos nomes dos autores, algumas bases utilizam iniciais com pontos e outras sem pontos.

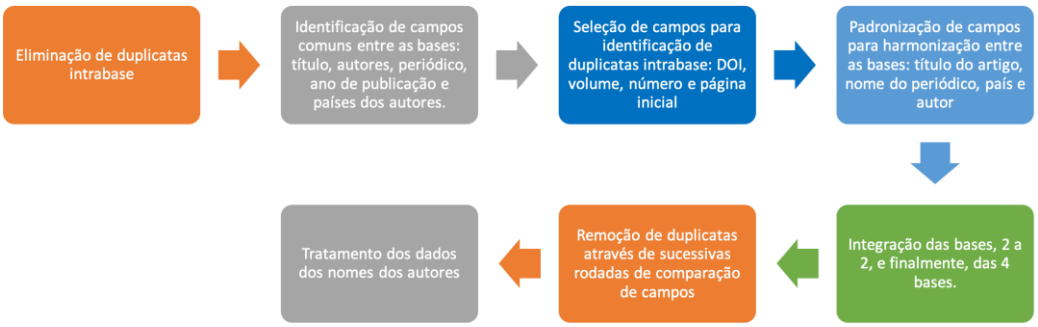
Tendo sido concluída a avaliação dos campos disponíveis, realizou-se então a integração das bases de dados, duas a duas. Inicialmente foram integradas a WoS e a SciELO, por terem a mesma estrutura, seguidas de PubMed e Scopus. Os dois conjuntos gerados foram então integrados de forma a obter um conjunto único com os registros das quatro bases de dados utilizadas.

Para identificar as sobreposições entre as bases e eliminar duplicatas, foram realizadas sucessivas rodadas de comparação de campos, inicialmente através do código DOI. Como esse dado não estava presente em todos os registros, foi necessária uma segunda etapa em que se utilizou como critério que fossem exatamente iguais: o título do artigo, o ano de publicação, o nome, o número e o volume do periódico. Uma última verificação foi feita ainda por

meio da análise individual dos títulos de artigos iguais para a eliminação de registros duplicados.

Por fim, foi realizado o tratamento dos dados referentes aos nomes dos autores, visando a comparação entre a posição dos principais autores considerando cada base individualmente e as quatro bases integradas. Para tanto, foi feito o uso de ferramentas de inteligência artificial do software *VantagePoint* para uniformizar o nome dos autores em cada base individualmente, com a conferência manual dos top 20 autores, verificando se haveria alguma grafia não identificada automaticamente. Nessa análise, a base SciELO foi desconsiderada devido à sua baixa representatividade no conjunto total. A Figura 1 sumariza as etapas da metodologia.

Figura 1 - Etapas metodológicas



Fonte: Elaborado pelos autores (2018).

2.3 Análise dos dados

O objetivo deste trabalho não foi a análise da produção na área, e sim a discussão sobre a integração dos dados. No entanto, o conjunto de registros de artigos únicos gerados permitiu o estudo de três variáveis presentes nas quatro bases e passíveis de análise: ano de publicação, autor do artigo e país de coautoria.

Para a análise das redes de coautoria de países foi realizada a exportação dos dados para o *VOSviewer* (versão 1.6.7), um software para construção e visualização de mapas bibliométricos. Os mapas criados a partir do *VOSviewer* permitem a análise gráfica baseada na coocorrência de itens analisados.

3 Resultados: apresentação e discussão

Para fins didáticos, os resultados são apresentados em três seções: uma relativa à coleta de dados e integração das fontes, a segunda sobre a padronização dos campos e remoção de duplicatas e, por fim, as análises feitas no conjunto de artigos único sobre Zika vírus publicados por pelo menos um autor brasileiro no período de 2014 a 2018.

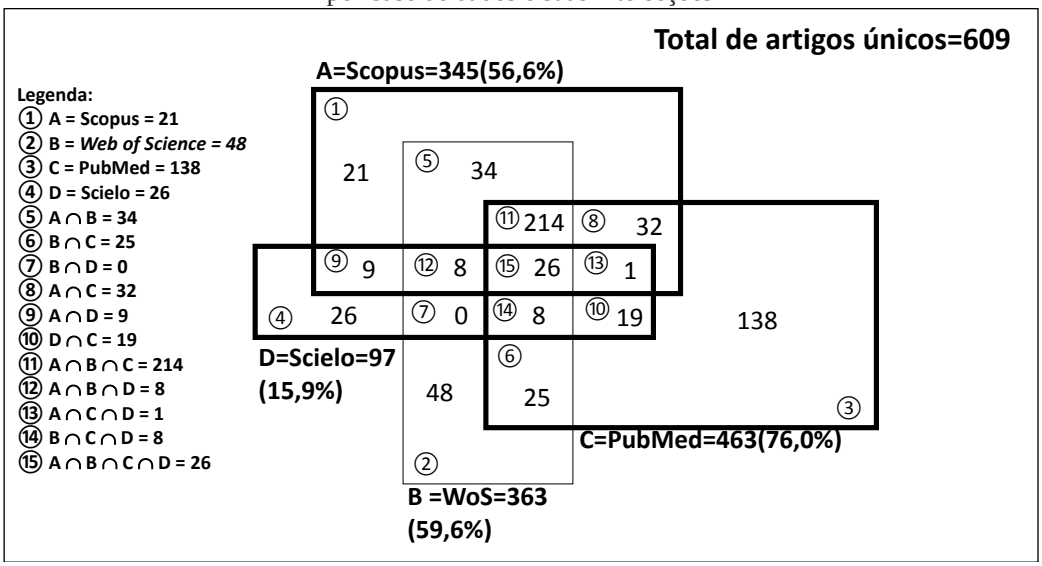
3.1 Coleta de dados e integração das fontes

A busca foi realizada em janeiro de 2018. Os dados resultantes da busca em cada base foram recuperados e importados para o software *VantagePoint*. A primeira etapa do tratamento de dados foi a verificação da existência de duplicatas intrabase. Enquanto nas bases WoS, SciELO e Scopus não foram identificadas duplicatas intrabase, no PubMed foram identificados apenas dois registros duplicados.

Em seguida, os conjuntos de cada base foram integrados dando origem a um único arquivo contendo os dados das quatro bases, com um total de 1.268 itens, que continham referências de artigos em duplicata. Nesse arquivo foi possível verificar as sobreposições de artigos nas quatro bases pesquisadas.

Na PubMed foram localizados 463 artigos, na WoS, 363, na Scopus, 345 e na SciELO, 97 artigos. No entanto, observa-se alguma sobreposição entre esses documentos, conforme pode ser visto no Diagrama de Venn apresentado na Figura 2. Esse diagrama mostra as interseções entre as combinações de duas bases, três bases e as quatro bases de dados utilizadas neste estudo, de forma a facilitar a compreensão no que se refere às vantagens de se proceder a integração de dados.

Figura 2 - Número de artigos sobre Zika vírus com um autor brasileiro (2014-2018) por base de dados e suas interseções



Fonte: Elaborado pelos autores (2018).

Em uma análise bibliometrica "padrão", em que se usaria, por exemplo, somente a WoS, haveria a perda de cerca de 40% de artigos na análise. No caso de uso da Scopus, a perda seria de cerca de 43%. Mesmo se fosse utilizada uma base específica do tema de estudo, a PubMed, ainda haveria a falta de 24% de documentos na temática. Se fosse usada a Scopus integrada com a WoS teríamos 426 documentos, o que representa um percentual de cerca de 70% do conjunto total[2]. No entanto, a união da WoS com a PubMed traria um resultado mais satisfatório, atingindo 91% dos resultados[3].

A Figura 2 também nos permite perceber que somente 26 artigos estão em todas as bases (A ∩ B ∩ C ∩ D), bem como a existência de 214 documentos que parecem ser mais relevantes por estarem em três bases, Scopus, WoS e PubMed (A ∩ B ∩ C). Não foi observado nenhum artigo presente concomitantemente nas bases WoS e SciELO (B ∩ D), o que nos faz refletir sobre a representatividade da produção nacional e local nas bases internacionais. Aparentemente, a WoS e a SciELO estão utilizando apenas a mesma interface de busca, pois os registros recuperados indicam que não há efetiva integração entre as coleções.

3.2 Padronização de campos e remoção de duplicatas

Para a correta eliminação de duplicatas, foi preciso inicialmente padronizar alguns campos no conjunto individual de cada base de dados como, por exemplo, título do artigo, país do autor e autor. O título dos artigos foi necessário uniformizar, pois no PubMed ele é finalizado com um ponto final, enquanto que nas demais bases esse modelo de padronização não é adotado. Outro problema encontrado em relação aos títulos é o uso de símbolos como α e β ou uso desses por extenso (alpha, beta).

Após a padronização inicial em cada conjunto individual, foi realizada a integração das bases de dados e a avaliação inicial do conjunto gerado conforme apresentado anteriormente. Em seguida, foi utilizado o recurso de remoção registros duplicados de forma que o conjunto final contivesse registros únicos de cada artigo identificado.

A primeira etapa de remoção de duplicatas foi realizada considerando o DOI: artigos com mesmo DOI são considerados repetidos. Esse campo estava presente em 97% dos registros. O arquivo continha 1.268 artigos, com a remoção usando o DOI, foi reduzido para 638 artigos.

Como se identificou que ainda havia títulos de artigos repetidos passou-se para a segunda fase de remoção de duplicatas, ainda com uma abordagem conservadora, ou seja, utilizando campos com menor potencial de falha de padronização. Nessa fase, considerou-se duplicata artigos que continham o título, o ano de publicação, o nome, o número e o volume do periódico iguais. Nessa fase, foram eliminadas mais 29 duplicatas, gerando um arquivo com 609 artigos.

Em seguida, tendo sido consolidado o conjunto final de artigos únicos, foi realizada então uma nova etapa de padronização de dados, sobretudo nos campos referentes ao país e nome do autor. Cabe notar que a padronização gera uma diminuição no número de itens relativo a cada campo. Se havia, por exemplo, no campo país, um artigo com Brazil e outro com Brasil, os dois artigos eram mantidos, com o campo uniformizado para Brazil. Por isso, a etapa de padronização visa reduzir o número de itens de cada campo (ao invés de Brasil e Brazil apenas Brazil).

O campo autor é o que apresenta maior dificuldade de padronização, devido à variedade de formas de abreviatura e uso ou não de pontos após as iniciais. Havia 5.281 nomes sem padronização, resultando em 4.501 autores com nomes uniformizados.

O campo país do autor está padronizado apenas na WoS, sendo necessária a verificação nos registros oriundos das outras bases. É importante observar também que, na SciELO, os nomes dos países estão em português, exigindo cuidado na padronização. Originalmente, havia 237 diferentes dados no campo país, e após a padronização restaram somente 107 países.

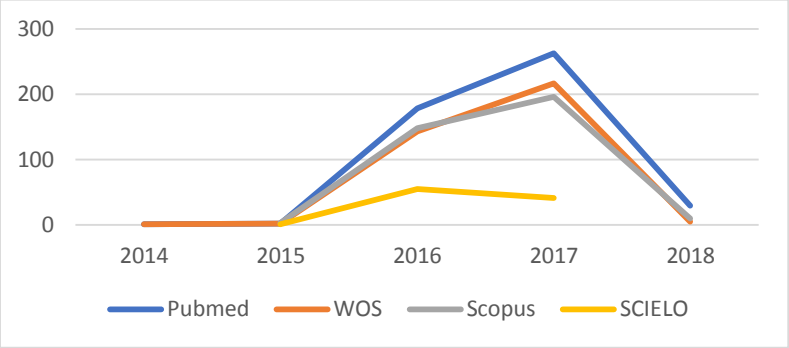
Destaca-se ainda que o uso de recursos de inteligência artificial através de processamento de linguagem natural, presentes no *VantagePoint*, foram fundamentais para a padronização dos campos, sendo, no entanto, realizada verificação de pertinência da padronização em cada uma das etapas.

3.3 Análise da produção científica sobre o Zika vírus

A integração de bases impõe a padronização dos dados, o que limita os campos que podem ser analisados. No entanto, há uma tendência de as bases uniformizarem seus dados. Além disso, a sofisticação das ferramentas de linguagem natural permite o uso de mecanismos que facilitam o tratamento de dados e possibilitam aproximações para análises bibliométricas confiáveis. Ainda que o estudo detalhado da produção na área fuja do objetivo central deste trabalho, algumas análises interessantes podem ser realizadas a partir do conjunto de dados gerado, as quais são apresentadas a seguir.

No que se refere ao ano de publicação, os registros indicam um crescimento expressivo a partir de 2015, ano no qual se iniciou uma epidemia de Zika vírus, sobretudo no Brasil, e a OMS declarou a doença uma emergência de saúde global. Como pode ser observado na Figura 3, o número de publicações apresenta um pico em 2017 em três das quatro bases estudadas[4].

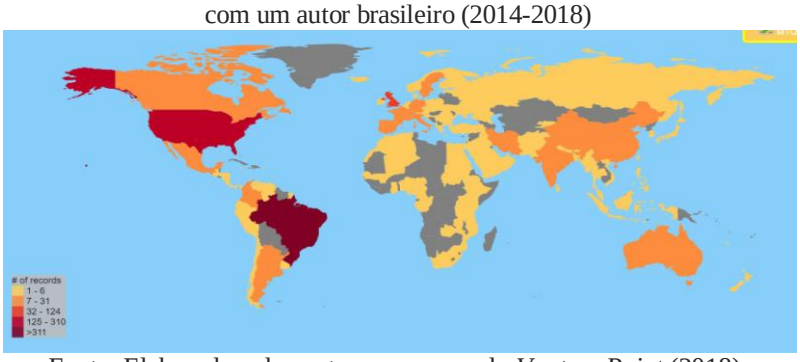
Figura 3 - Número de artigos sobre Zika vírus com um autor brasileiro, distribuído por ano e por base de dados



Fonte: Elaborado pelos autores (2018).

A análise dos países dos coautores indica que não há diferença significativa entre os dados coletados para esse campo em cada uma das bases de dados. Assim, a Figura 4 apresenta o mapa global, cuja intensidade da cor indica maior colaboração. Como o estudo foi focado em publicações com pelo menos um autor brasileiro, o Brasil é o país com o vermelho mais forte no mapa. A observação da Figura 4 indica uma amplitude de colaboração nos diversos continentes, em que se destaca a colaboração no eixo Sul, onde há maior incidência de doenças como a Zika.

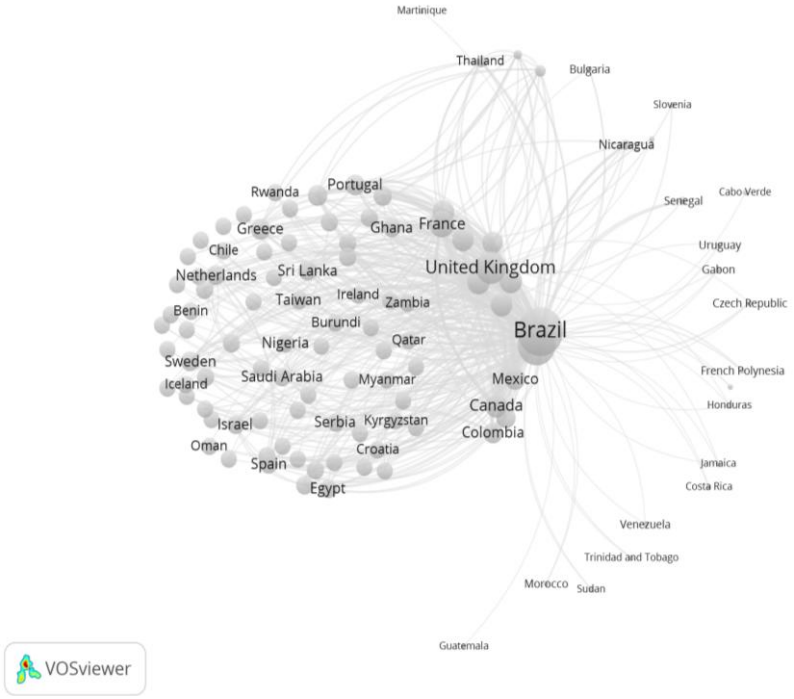
Figura 4 - Mapa global de coautorias de artigos sobre Zika vírus com um autor brasileiro (2014-2018)



Fonte: Elaborado pelos autores com uso do *VantagePoint* (2018).

Outro aspecto relevante que pode ser observado é a quantidade de países com que os autores brasileiros têm parceria. A Figura 5 apresenta o mapa de rede de coautoria: a força do *link* entre países indica o número de publicações conjuntas, enquanto o tamanho dos "nós" referentes a cada país representa o seu volume de publicações. A análise de coautoria incluiu informações de todos os 107 países associados às instituições dos 4.501 autores identificados no estudo.

Figura 5 - Rede de países de coautoria de artigos sobre Zika vírus com um autor brasileiro (2014-2018)



Fonte: Elaborado pelos autores com uso do *VOSViewer* (2018).

Os dados de nomes de autores foram analisados para comparar a posição em que eles se encontram em cada base individualmente e no conjunto integrado das quatro bases. O Quadro 1 apresenta a lista dos autores com dez ou mais publicações indexadas na WoS e a posição deles nesse conjunto. As colunas seguintes apresentam o posicionamento desses autores por número de publicações nas bases de dados Scopus e PubMed, bem como o posicionamento considerando as quatro bases integradas.

Quadro 1 - Comparativo da posição dos autores mais produtivos dentre as bases do estudo de artigos sobre Zika vírus com um autor brasileiro (2014-2018)

Autor (número de artigos na WoS)	Posição na base			
	WoS	Scopus	PubMed	4 bases
Cordeiro, Marli Tenorio	1	1	1	1
Bispo de Filippis, Ana Maria	2	3	2	4
van der Linden, Vanessa	2	2	5	5
Brasil, Patricia	3	4	2	3
Nogueira, Mauricio Lacerda	4	8	4	6
Tanuri, Amilcar	5	6	3	7
Vasconcelos, Pedro F C	5	5	2	2
de Oliveira, Wanderson Kleber	6	9	12	11
Lourenco-de-Oliveira, Ricardo	7	8	7	9

Fonte: Elaborado pelos autores (2018).

Observa-se que há uma variação no posicionamento dos autores nas diferentes bases de dados. Por exemplo, o autor que tem a segunda maior produção, considerando as quatro bases de dados, está em 5° tanto na WoS como na Scopus. Em estudos de indicadores para financiamento de pesquisas, é usual se considerar apenas uma base de dados, o que pode trazer uma repercussão no investimento nas pesquisas de um autor com uma produtividade maior do que a representada em uma única base. Por outro lado, um autor que ocupa, por exemplo, o 2º lugar na WoS, quando considerada as quatro bases integradas, ele acaba ocupando o 5° lugar.

4. Considerações finais

A investigação apresentada indica que o uso de uma única fonte, por melhor que seja sua cobertura, apresenta uma perda de 24 a 43% de artigos, para o estudo do Zika vírus. A integração de pelo menos duas fontes diferentes, reduziria para nove a 30% a perda de documentos recuperados.

A complementariedade das fontes informacionais pode ser percebida pelo baixo número de artigos comum entre elas (4%) e presentes em pelo menos três bases de dados (35%).

Este estudo indica que é necessário estimular ações para atender as recomendações do Manifesto de Leiden (HICKS, 2015). Dentre as dez recomendações sugeridas para orientar o uso de métricas em avaliação da pesquisa, a terceira refere-se à importância de proteger a excelência da pesquisa desenvolvida e publicada localmente. Em geral, os estudos bibliométricos tomam como base as publicações em inglês, mas em muitas áreas, como a Saúde Pública, por exemplo, há inúmeras pesquisas que tem dimensão regional ou nacional. O pluralismo da sociedade tende a ser suprimido devido a necessidade da elaboração de artigos de alto impacto para periódicos de língua inglesa. Métricas desenvolvidas a partir de literatura de alta qualidade em outros idiomas são úteis para identificar e premiar a excelência na pesquisa relevante localmente. O uso de diferentes bases de dados tende a minimizar essa questão, portanto a integração das bases e novas metodologias desenvolvidas nesse sentido são caminhos importantes a serem investigados.

Referências

ALENCAR, Maria Simone de Menezes; BOCHNER, Rosany; GIACOMETTI, Daniel. A importância da integração de dados: a produção científica nacional em Zika. In: ENCONTRO BRASILEIRO DE BIBLIOMETRIA E CIENTOMETRIA, 6., 2018, Rio de Janeiro. **Anais...** Rio de Janeiro: UFRJ, 2018.

ARAÚJO, Kizi Mendonça; SILVA, Cícera Henrique; GUIMARÃES, Maria Cristina Soares. Produção científica e doenças emergentes: o caso da Zika. In: ENCONTRO NACIONAL DE PESQUISA EM CIÊNCIA DA INFORMAÇÃO,17., 2016, Salvador. **Anais...** Salvador: ANCIB, 2016.

ARAÚJO, Kizi Mendonça; SILVA, Cícera Henrique; GUIMARÃES, Maria Cristina Soares; LINS, Rosane Abdala; ASSEF NETO, Rosangela Cordeiro de Souza. A produção científica sobre zika em periódicos de acesso aberto. **RECIIS** – Rev. Eletron. Comun. Inf. Inov. Saúde, v. 11, n. supl., out. /dez. 2017.

ARCHAMBAULT, Éric et al. Comparing bibliometric statistics obtained from the Web of Science and Scopus. **Journal of the American society for information science and technology**, v. 60, n. 7, p. 1320-1326, 2009.

BUENO, Flávia Thedim Costa; GARCÍA, Mónica; MOYA, José; LÖWY, Ilana; BENCHIMOL, Jaime L.; CERQUEIRA, Roberta C.; CUETO, Marcos. Zika e Aedes aegypti: antigos e novos desafios. **História, Ciências, Saúde – Manguinhos**, v. 24, n. 4, p.1161-1179, out./ dez. 2017.

BURNHAM, Judy F. Scopus database: a review. **Biomedical digital libraries**, v. 3, n. 1, p. 1, 2006.

FALAGAS, Matthew E. et al. Comparison of PubMed, Scopus, web of science, and Google scholar: strengths and weaknesses. **The FASEB journal**, v. 22, n. 2, p. 338-342, 2008.

GAVEL, Ylva; ISELID, Lars. Web of Science and Scopus: a journal title overlap study. **Online information review**, v. 32, n. 1, p. 8-21, 2008.

GHENO, Ediane Maria; SCHÜLER-FACCINI, Lavinia; SOUZA, Diogo Onofre; CALABRÓ, Luciana. Panorama e características da produção científica brasileira e internacional sobre Zika vírus. In: ENCONTRO BRASILEIRO DE BIBLIOMETRIA E CIENTOMETRIA, 5., 2016, São Paulo. **Anais...** São Paulo: USP, 2016. p. A26.

GHENO, Ediane Maria; SCHÜLER-FACCINI, Lavinia; SOUZA, Diogo Onofre; CALABRÓ, Luciana. Vírus Zika: produção e colaboração científica brasileira. In: REUNIÃO ANUAL DA SBPC, 69., 2017, Belo Horizonte. **Anais...** Belo Horizonte: UFMG, 2017. p. 1-4.

HARZING, Anne-Wil; ALAKANGAS, Satu. Google Scholar, Scopus and the Web of Science: a longitudinal and cross-disciplinary comparison. **Scientometrics**, v. 106, n. 2, p. 787-804, 2016.

HICKS, Diana et al. The Leiden Manifesto for research metrics. **Nature**, v. 520, n. 7548, p. 429, 2015.

JACSO, Peter. As we may search: comparison of major features of the Web of Science, Scopus, and Google Scholar citation-based and citation-enhanced databases. **Current science,** v. 89, n. 9, p. 1537-1547, 2005.

MOED, Henk F.; MARKUSOVA, Valentina; AKOEV, Mark. Trends in Russian research output indexed in Scopus and Web of Science. **Scientometrics**, v.116, n.2, p. 1-28, 2018.

MONGEON, Philippe; PAUL-HUS, Adèle. The journal coverage of Web of Science and Scopus: a comparative analysis. **Scientometrics**, v. 106, n. 1, p. 213-228, 2016.

MUELLER, Suzana Pinheiro Machado. Estudos métricos da informação em ciência e tecnologia no Brasil realizados sobre a unidade de análise artigos de periódicos. **Liinc em Revista**, Rio de Janeiro, v. 9, n. 1, p. 6-27, maio 2013.

NEUHAUS, Christoph; DANIEL, Hans-Dieter. Data sources for performing citation analysis: an overview. **Journal of Documentation**, v. 64, n. 2, p. 193-210, 2008.

NORRIS, Michael; OPPENHEIM, Charles. Comparing alternatives to the Web of Science for coverage of the social sciences' literature. **Journal of informetrics**, v. 1, n. 2, p. 161-169, 2007.

NUNES, Magda Lahorgue; CARLINI, Celia Regina; MARINOWIC, Daniel; KALIL NETO, Felipe; FIORI, Humberto Holmer; SCOTTA, Marcelo Comerlato; ZANELLA, Pedro Luis Ávila; SODER, Ricardo Bernardi; COSTA, Jaderson Costa. Microcephaly and Zika virus: a clinical and epidemiological analysis of the current outbreak in Brazil. **Jornal de Pediatria**, v. 92, n. 3, p. 230-240. 2016

RIBEIRO, Kíssila da Conceição; PAULO, Ana Carolina Laurindo; DANTIER, Rui Manoel Pinto; BATISTA, Fabio Barbosa; MIRANDA, Guilherme Melo. Zika vírus: indicadores bibliométricos das publicações na base Scopus. In: CONGRESSO DE INTERDISCIPLINARIDADE DO NORDESTE FLUMINENSE, 1., 2016, Itaperuna. **Anais...** Itaperuna: Instituto Federal Fluminense, 2016.

VAN LEEUWEN, Thed N. et al. Language biases in the coverage of the Science Citation Index and its consequencesfor international comparisons of national research performance. **Scientometrics**, v. 51, n. 1, p. 335-346, 2001.

Database integration in bibliometric studies: the Brazilian scientific production in Zika

Abstract: Based on the case study of the Brazilian scientific production involving Zika virus, this article aims to discuss the importance of integrating different data sources to better represent the landscape of the research, highlighting how the restriction of sources impacts the results of a bibliometric analysis. Four scientific databases of peer-reviewed literature were used to identify the scientific production of Brazilian authors related to Zika virus: Web of Science and Scopus as two general sources; PubMed as a specific database related to the field of the study; and SciELO as a database with a more regional focus. The analysis of the results indicates the relevance of integrating different databases in bibliometric studies in order to mitigate distortions and provide a consistent vision of the scientific research landscape in a certain field.

Keywords: Bibliometrics. Databases. Zika. Information Sources.

Recebido:24/09/2018
Aceito: 13/12/2018

[1] Software comercial, desenvolvido pela Georgia Tech e comercializado pela Search Technology. É uma ferramenta de mineração de texto para descoberta de conhecimento em resultados de busca em bases de dados estruturadas tais como de artigos científicos e de patentes.
[2] Scopus + Web of Science → $(A \cup B) = A + B - (A \cap B) - (A \cap B \cap C) - (A \cap B \cap D) - (A \cap B \cap C \cap D) = 426$ → 426/609 = 69,95%
[3] Web of Science + PubMed → $(B \cup C) = 553$ → 553/609 = 90,80%
[4] Para elaboração deste gráfico foi considerado o conjunto de registros com duplicata, pois o objetivo era identificar a cobertura da base em relação ao ano de publicação.